Building new knowledge from distributed scientific corpus

HERBADROP & EUROPEANA

Two concrete case studies for exploring big archival data

Pascal Dugenie* Nuno Freire† Daan Broeder‡
*CINES, Centre Informatique National de l'Enseignement Superieur, Montpellier, France, dugenie@cines.fr
†INESC-ID/Europeana DSI, Den Haag, NL, nuno.freire@tecnico.ulisboa.pt
‡MEERTENS Institut (Afdeling Technische Ontwikkeling), Amsterdam, NL, daan.broeder@meertens.knaw.nl

***Abstract*—This paper presents approaches for building new knowledge using emerging methods and big data technologies together with archival practices.**

Two cases studies have been considered. The first one called *HERBADROP* is concerned with preservation and analysis of herbarium images. The second one called *EUROPEANA* investigates how to facilitate the re-use of cultural heritage language resources for research purposes. The common point between these two case studies is that they are both concerned with the use of valuable heritage resources within the EUDAT (European Data) infrastructure. *HERBADROP* leverages on the data services provided by EUDAT for long-term preservation, while *EUROPEANA* leverages on EUDAT to achieve citability and persistent identification of cultural heritage datasets.

EUDAT[1] is an initiative of some of the main European data centers and together with community research infrastructure organisations, to build a common eInfrastructure for general research data management.

In this paper, we show how technologcal trends may offer some new research potential in the domain of computational archival science in particular appraising the challenges of producing quality, meaning, knowledge and value from quantity, tracing data and analytic provenance across complex big data platforms and knowledge production ecosystems.

I. Introduction

A. State-of-the-art

UNTIL recently, in most research domains, a massive effort was demanded to build new knowledge because scientific corpora were not easily accessible. With the large digital archiving programs, the extraction of knowledge from scientific collections is becoming a fast growing activity [15].

Various approaches for data mining and knowledge building are proposed in the literature. However, the specificities of every data collections may result to quite heterogeneous analysis models. For instance, a probabilistic method that would be suitable for a category of data would reveal to be ineffective in another domain.

Several attemps for proposing probabilistic approach or using topic modelling to connected archival data have been published recently [21]. As pointed in [20], digitisation initiatives for archives have created huge textual corpora but they are often of bad OCR quality and lack of metadata.

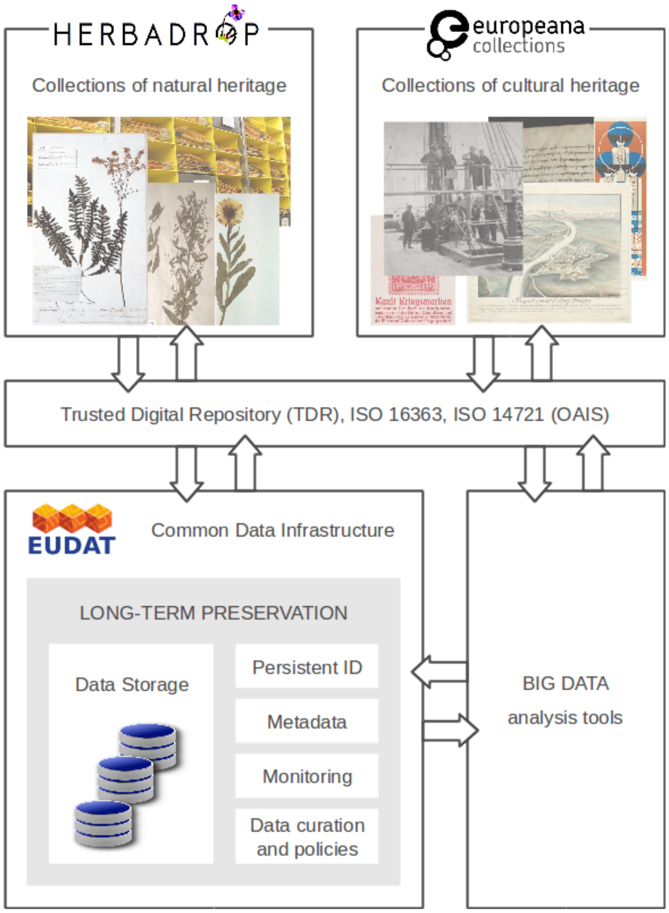

HERBADROP
Collections of natural heritage
europeana collections
Collections of cultural heritage
Trusted Digital Repository (TDR), ISO 16363, ISO 14721 (OAIS)
EUDAT
Common Data Infrastructure
LONG-TERM PRESERVATION
Data Storage
Persistent ID
Metadata
Monitoring
Data curation and policies
BIG DATA analysis tools

Fig. 1. A layered view of provision of services and facilities to preserve and exploit cultural and natural heritage data.

Other approaches [19], [25] propose to explore identity and provenance of data. Data provenance is a very fundamental aspect for establishing trust in the data being produced [18]. For instance, Xu et al. [25] is exploring corpora using comparison algoritms based on 165 millions of records to compare degree of similarity of records are presented in a match score distribution histogram. These results show that comparisons may help curators to understand the reasons for differences and similarities.

However in any of these approaches, data curation requires

[1] EUDAT initiative is funded by a series of European projects, wherof the latest EUDAT2020 that is also funding the Pilot projects working on the described use-cases. See http://eudat.eu

more automated scalable infrastructure and computation services as shown in Fig. 1. such as the ones offered by a Trusted Digital Repository (TDR) over a EUDAT infrastructure and linked with a set of big data analysis tools. TDR is a concept than is obtained with strict archival practices and standardisation (see [9], [17], [18]).

B. About EUDAT

EUDAT [12] is a European e-infrastructure of integrated data services and resources in support of research. This infrastructure and its services have been developed in close collaboration with over 50 research communities spanning across many different scientific disciplines, with more than 20 major European research organizations, data centres and computing centres involved. EUDAT CDI has emerged as a result of two subsequent FP7 and Horizon 2020 projects, with the actual services focused on different aspects of data management and data use and supported by a variety of information technology stacks. Researchers can use EUDAT data services to support their scientific investigations and data management. By joining the EUDAT CDI, research organizations are able to better serve their users across borders and better support cross-national collaborations. The EUDAT initiative is funded by a series of European projects, whereof the latest EUDAT2020 that is also funding the Pilot projects working on the described use-cases. The EUDAT projects support realising the EUDAT vision to build a CDI which is a collaboration of large data and compute centers, generic digital solution providers for research domain and thematic (discipline specific) service and data providers. Part of the CDI is also a portfolio of general research data management services the are being continuously further developed and tested in different collaborations and projects with research community organisations and project groups.

The EUDAT service portfolio includes various services for data management:

- B2SHARE: service for managing sharing of small-scale and long tail data
- B2SAFE: service for robust, safe and highly available replication service for managing large-scale data in community and departmental repositories
- B2STAGE: service for managing data transfers between EUDAT storage and high-performance computing
- B2FIND: service for data discovery
- B2DROP: service for secure and trusted data exchange and sharing

In addition, all these EUDAT services are integrated using common services:

- B2HANDLE: service for managing identification using Persistent IDentifiers (PID)
- B2ACCESS: service for managing Authorization and Authentication (AAI)

Historically, EUDAT services have been built with only a few considerations for conscious data curation, with secure and controlled access to data being one of the major initial goals to achieve. Other aspects of data curation started playing a more prominent role when services matured to production stage and became a part of an operational collaborative infrastructure [4].

Currently there are collaboration projects with communities over a wide range of disciplines from the Life Sciences, Humanities, Earth Sciences and Physics. A large number of these collaborations was started by a call for collaborations that the EUDAT2020 project initiated in 2015 and resulted in 24 so named Data Pilots that together provide collaborations and test-beds with a huge variety in disciplines, community project group sizes and IT maturity chalenges.

The EUDAT engagements with the Data Pilots has advanced greatly the EUDAT's comprehension of the needs of the research communities and projects wrt. data management. The pilots have been instrumental in steering EUDAT's development agenda and cementing contacts with a large number of research communities until then outside EUDAT's immediate circle partner communities. At the same time the Pilots have profited considerably from the collaboration with EUDAT's delivering ressources and consultancy with respect to creating and using infrastructure for research.

The two use-cases described in this paper are excellent examples of that variety and show how the different services are used to solve community data management needs.

II. HERBADROP USE-CASE

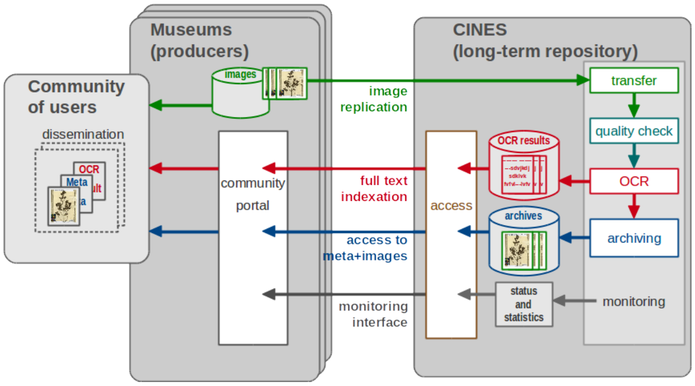


Fig. 2. Herbadrop architecture and processes

A. Background

For several centuries, Natural History Collection (NHC) institutes (i.e. museums and botanical garden) accross the world have been responsible for preserving the physical copies of herbaria. These herbaria are collections of plants sticked on a sheet with annotations that describe a given specimen such as the one represented in Fig. 3. Each herbarium specimen has been collected, carefully prepared and annotated by botanists.

B. Challenges of large-scale digitization of herbaria archives

Herbaria hold large numbers of collections: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. Through the years, these collections have

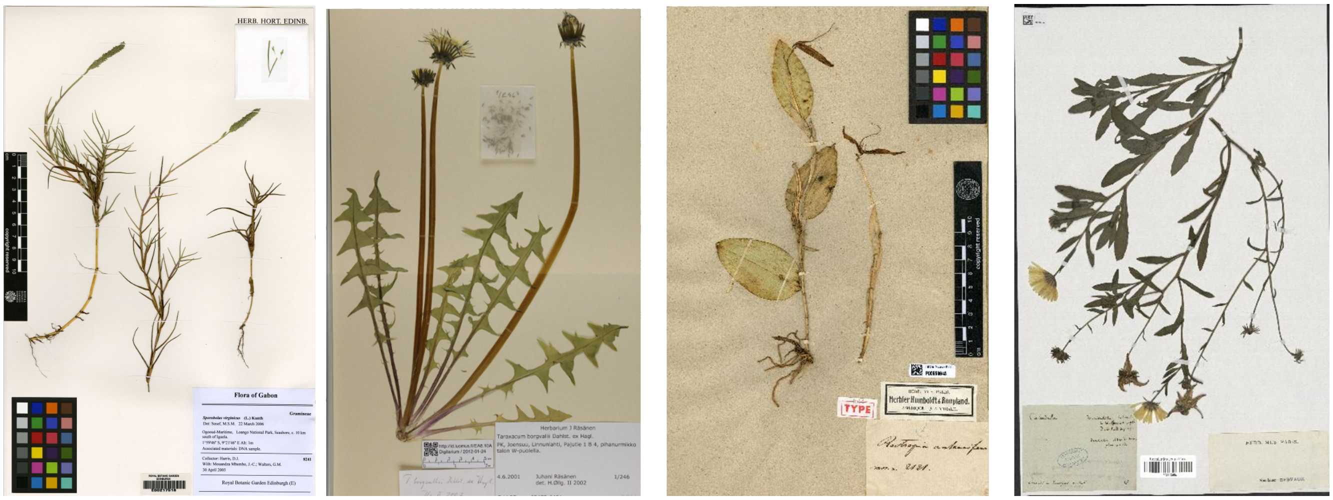


Fig. 3. Representation of typical herbarium specimen: flora of Gabon, a taxacum, a restrepia antennifera and a calendula denticulata

been studied and enriched by taxonomists [1]. This represents altogether a very precious yet fragile scientific basement.

Therefore, in the digital era, it has become obvious for the NHC institutes to yield a vaste digitization campaign of their herbaria and limit the manipulation of the physical copies [16], [22], [3], [13], [2].

Nevertheless, the responsibility of preserving the the digital versions is a new challenge. High-resolution images of herbarium specimens require substantial bandwidth and disk space. Moreover, ensuring long-term preservation of digital objects is not a straight forward task for organisms that cannot afford to manage high volumes and acquire the suitable knowledge on data format that will still be readable in many years.

Another significant challenge is the new perspective for performing all kind of image analysis using intensive computing and post-processing techniques. Again, this requires computing skills that are not always obvious in NHC institutes.

Since data storage and image processing are not natural skills of NHC institute, some of them decided to rely on a third party that can provide a Trusted Digital Repository (TDR) and a shared access to the data for the whole community. Thus, all demanding tasks in terms of operational constrains, respect of formal OAIS processes [17], etc., are delegated.

C. About the HERBADROP data pilot

Fortunately, since 2012, the EUDAT project has facilitated the emergence of scientific collaborations between communities and data centres and merges domain skill with computing specific skills. One of the EUDAT data pilot dedicated for this purpose is called Herbadrop [11], [14] is one of the two use-cases presented in this paper.

1) Objectives: The two core objectives of the Herbadrop data pilot are:

- long-term preservation of scientific natural heritage: collections of digitalized herbaria are transferred from several European museums and botanical gardens to a TDR.
- extraction of written information from these images by using Optical Character Recognition (OCR) analysis using intensive computing.

The long term preservation is an ongoing task that is mainly focusing on operational aspects. Therefore, for the purpose of this paper, the focus in made on reporting results for the analysis of the OCR results.

2) Community partners: Initially, the consortium was formed by five NHC institutes from Finland, France, Germany, Netherlands and Scotland. Their common objective was to share their herbaria for future research projects by making the specimen images and data available on-line from different institutes allows cross domain research and data analysis for botanists and researchers with diverse interests (e.g. ecology, social and cultural history, climate change).

BGBM (De.): The Botanischer Garten und Botanisches Museum (BGBM) of Berlin is to a large extent based on its scientific plant collections. A central element of its activities is taxonomic research, through which plants are identified, described, named and classified.

MNHN (Fr.): The Musum National d'Histoire Naturelle (MNHN) of Paris is in charge of the main collection of botanical and zoological specimens in France. Between 2008 and 2012, it completed a massive digitization program of the herbarium specimens, putting online nearly 6 millions of images. It will greatly benefit of the pilot for both long-term preservation of image files and extraction of the label information.

RBGE (Sco.): The Royal Botanic Garden Edinburgh (RBGE) has a very active herbarium of 3 million specimens and living collection of around 64,000 plants. All of the living collection records, including more than 40,000 linked images, are online and 300,000 of the herbarium specimens are images at high resolution which are available online. RBGE has incorporated OCR technology into the digitisation workflow and is currently testing Handwritten Text Recognition.

Digitarium (Fin.): Digitarium is the digitisation centre of the Finnish Museum of Natural History and the University of Eastern Finland. In 2014, Digitarium coordinated a H2020 proposal for designing a European distributed digitisation infrastructure for natural heritage (acronym: ICEDIG).

Naturalis (NL): Naturalis Biodiversity Center (Naturalis, Leiden, The Netherlands) is the merger of the National Museum of Natural History, the Zoological Museum of Amsterdam and the National Herbarium of the Netherlands. Naturalis has just finished its mass digitisation project in which 4,2 M higher plants were scanned, databased and published. Naturalis decided to not extend its membership after the fist phase of Herbadrop, but will come back since it is partner of the forthcoming ICEDIG project.

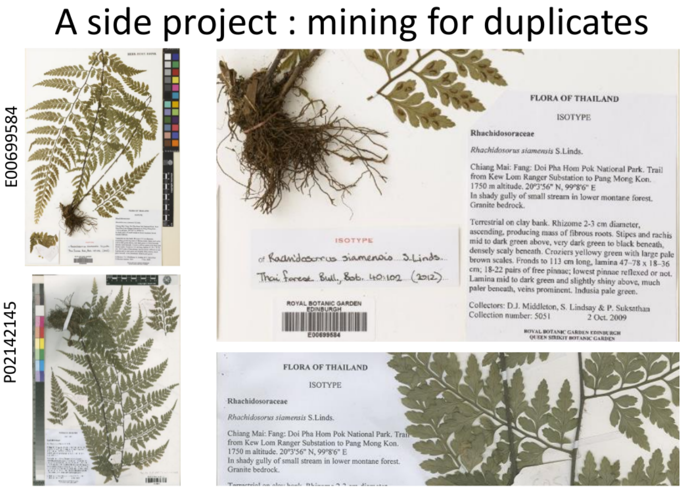
A new partner, the **Botanic Garden of MEISE**, Belgium, joined the consortium during 2017. The herbarium of Botanic Garden Meise houses around 4 million specimens. The Vascular Plant Herbarium contains three main collections: the General Herbarium with more than one million specimens; the Belgian Herbarium with about 200,000 specimens; and the African Herbarium comprising at least one million specimens (of which over half are from central Africa). The 800,000 specimens in the Cryptogam Herbarium consist of mosses, lichens, algae, fungi and myxomycetes.

3) EUDAT service provider: As EUDAT partner, the Centre Informatique National de l'Enseignement Superieur (CINES) has been identified to offer a TDR for long term archival and also as a suitable HPC centre.

Since herbarium collections are spread all over the world, a wide data infrastructure such as EUDAT in a European scale is required. EUDAT offers the suitable services for ingesting and processing the specimen images. CINES plays the role of entry point in the infrastructure. The B2SAFE running instance at CINES is used in the first step of the ingestion process. The ingestion into B2SAFE is carried out in accordance with the centralized persistent identifiers (PID) management system used in EUDAT (i.e. EPIC handle). Furthermore, the discovery, sharing and visualization of the data objects can be performed with the EUDAT B2FIND service.

D. *Expected domain legacy*

One important outcome from the Herbadrop data pilot is that the specimens will be discoverable by the entire scientific community. Thus, undescribed species stored in herbaria can be examined by experts to aid identification and discovery of new species. Distribution information for species over time can be evaluated and these data could provide evidence of the point of time when an invasive species first occurred in a certain area. Historians could analyze herbarium data to create itineraries for historical characters. The data can be used to calibrate predictive models of the oncoming changes in biodiversity patterns under global threats. This diverse information will be useful for a wide user community including conservationists, Policy makers, and politicians. Progess and status During the last eighteen months, project partners built an infrastructure for digitized herbarium specimens based on B2SAFE service and the CINES repository. The workflow uses the EUDAT B2SAFE function to transfer images. Data quality controls are performed like an integrity and antivirus check, a file format validation or a metadata control. Some of them are done on high performance computer as well as the OCR analysis. The OCR results are indexed in full text in a search engine. An access function provides access to images and OCR results with a full text search. At the end of the first phase of the pilot (July 2017), more than 4.5 Million images have been processed, equaling more than 27 TB of volume and 200 000 hours of computation power. The pilot extends to nine additional months (until February 2018) to add new treatments on images, supply consolidated metadata and finalize some functions of the long-term preservation repository (like tracking service or statistics). One of the objective is to obtain a DSA (Data Seal of Approval) agreement to demonstrate the quality of service.



A side project : mining for duplicates

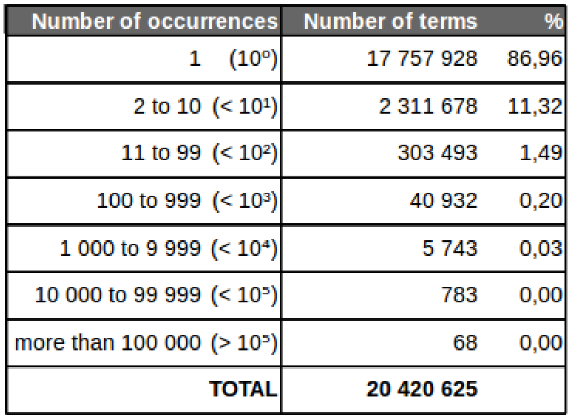
E00699584
P02142145
FLORA OF THAILAND
ISOTYPE
Rhachidosoraceae
Rhachidosorus siamensis S.Linds.
Chiang Mai: Fang: Doi Pha Hom Pok National Park. Trail from Kew Lom Ranger Substation to Pang Mong Kon. 1750 m altitude. 20°3'56" N, 99°8'6" E
In shady gully of small stream in lower montane forest. Granite bedrock.
Terrestrial on clay bank. Rhizome 2-3 cm diameter, ascending, producing mass of fibrous roots. Stipes and rachis mid to dark green above, very dark green to black beneath, densely scaly beneath. Croziers yellowy green with large pale brown scales. Fronds to 113 cm long, lamina 47–78 x 18–36 cm; 18-22 pairs of free pinnae; lowest pinnae reflexed or not. Lamina mid to dark green and slightly shiny above, much paler beneath, veins prominent. Indusia pale green.
Collectors: D.J. Middleton, S. Lindsay & P. Suksathan
Collection number: 5051 2 Oct. 2009
ROYAL BOTANIC GARDEN EDINBURGH
QUEEN SIRIKIT BOTANIC GARDEN
ISOTYPE
of Rachidosorus siamensis S.Linds.
Thai Forest Bull., Bot. 40:102 (2012)
ROYAL BOTANIC GARDEN EDINBURGH
E00699584
FLORA OF THAILAND
ISOTYPE
Rhachidosoraceae
Rhachidosorus siamensis S.Linds.
Chiang Mai: Fang: Doi Pha Hom Pok National Park. Trail from Kew Lom Ranger Substation to Pang Mong Kon. 1750 m altitude. 20°3'56" N, 99°8'6" E
In shady gully of small stream in lower montane forest. Granite bedrock.

Fig. 4. Illustration of possibilities to exploit OCR techniques to identify duplicates from different museums. These are two specimen of the same plant located in two different museum. The specimen above is located in RBGE, Edinburgh, identified with the barcode E00699584, whereas the specimen below is located in MNHN, Paris, identified with the barcode P02142145. The text written in the label show that these two specimen have been collected with similar information attached to it (presented in [5].

E. *A prelimilary overview of OCR analyses of herbarium specimen*

New methods of extracting information from the specimen labels have been developed using OCR but using this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, and the variable and ancient fonts [8]. Most of the information is only available only using handwritten text recognition or botanical pattern recognition which is less mature technology than OCR. There are numerous possibilities to exploit these OCR results, such as identifying specimen that are duplicated in different museums (see Fig. 4). Initial tests with OCR tools showed poor results [10]. However, we may expect that even noisy results can be exploitable on a large number of dataset. The scientific approach here is mainly statistical. At this stage of the project, the 4.5 million specimen sheets[2] analysed through OCR techniques come from four different European collections. The OCR tool used all along this preriod is Tesseract [23] using multiple dictionaries (lat, eng, fra, ger, esp).

[2] All sepcimen images are under Creative Common licence CC-BY. The intellectual property belongs to MNHN, BGBM, RBGE and Digitarium.



Number of occurrences	Number of terms	%
1 (10^0)	17 757 928	86,96
2 to 10 $(< 10^1)$	2 311 678	11,32
11 to 99 $(< 10^2)$	303 493	1,49
100 to 999 $(< 10^3)$	40 932	0,20
1 000 to 9 999 $(< 10^4)$	5 743	0,03
10 000 to 99 999 $(< 10^5)$	783	0,00
more than 100 000 $(> 10^5)$	68	0,00
TOTAL	20 420 625	

Fig. 5. Distribution of terms by number of occurrences

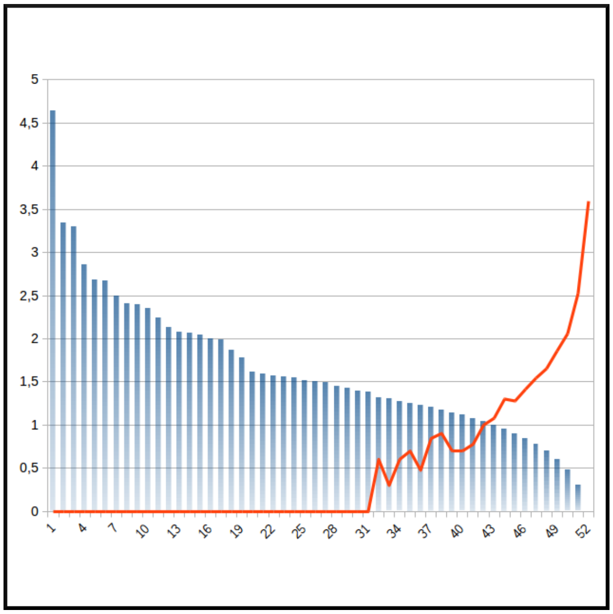


Fig. 6. OCR dispersion for the term vernaculaire (Log-Log scale). The number of occurence of the meanigful term is represented in lue bars on the left side. The red curve represents the level of noise (i.e. terms with a single occurrence are counted on the right side).

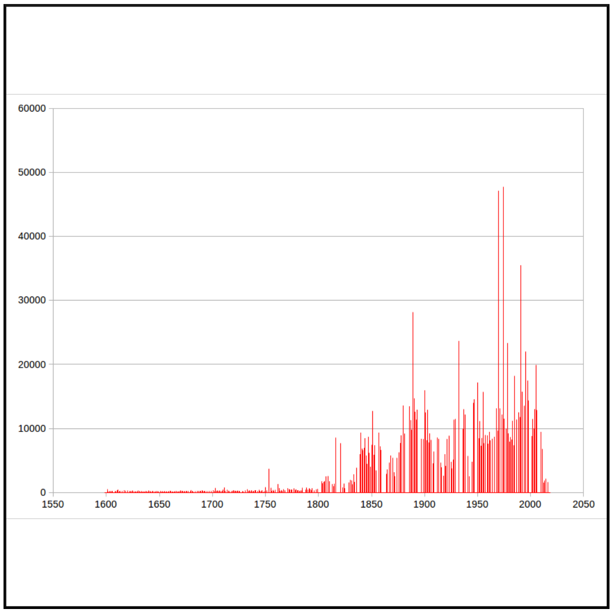
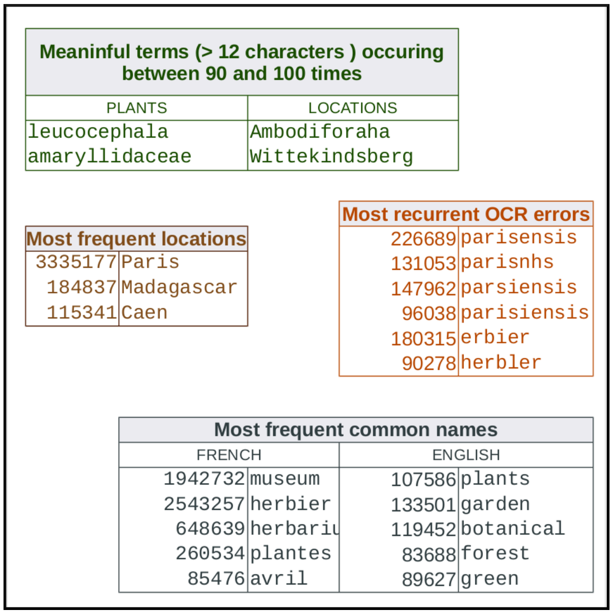


Fig. 7. Distribution of all dates identified in the labels



Meaninful terms (> 12 characters) occuring between 90 and 100 times	
PLANTS	LOCATIONS
leucocephala	Ambodiforaha
amaryllidaceae	Wittekindsberg

Most frequent locations	
3335177	Paris
184837	Madagascar
115341	Caen

Most recurrent OCR errors	
226689	parisensis
131053	parisnhs
147962	parsiensis
96038	parisiensis
180315	erbier
90278	herbler

Most frequent common names			
FRENCH		ENGLISH	
1942732	museum	107586	plants
2543257	herbier	133501	garden
648639	herbariu	119452	botanical
260534	plantes	83688	forest
85476	avril	89627	green

Fig. 8. Various statistics on terms and errors

F. Statistical analysis of the OCR results

The first thing to dertermine is the kind of knowledge we can extract from the OCR output.

Then, how to extract any knowledge from a corpus that comes from various places and contains very heterogeneous information and languages. The fact that the access has been made centralized rather than distributed, facilitate the performance of any kind of analyses.

1) Methodology: Since the analysis is purely statistical and nor domain oriented, the metodology consists to adapt functional programming techniques such as mapping, filtering and reducing using regexp based algorithms.

The objective of this first approach is reporting frequencies of occurrence of terms in order to evaluate OCR quality and possible semantic links between data objects (i.e specimen).

OCR results are extracted from the storage system, then converted into key-values pairs. The key correspond to the term, the value corresponds to the number of occurrences.

2) Observations: At this stage, there is no interpretation from domain specialists. This statistical analysis aims to provide a overview of what can be identified by aggregating large quantity of data containing many approximations. These results may serve as input for future crowd-sourcing services dedicated for curating the data and consolidating semantical links [24], [6]. The first observation when analysing the OCR results is its long-tail nature. There are more than 211 millions of occurrences and 20 millions of different terms identified. The distribution of the number of occurrences versus the number of terms follows a Pareto distribution with a very high k-factor. For instance, more than 85 % (17 millions out of 20 millions) of terms occur only once. Their probability of being noise (incorrect term) is very high. In the opposite, few hundred (0,004 %) of terms occur more than 10 000 times.

One of the characteristics of herbaria metadata is that they contains many seldom references which increase the long-tail nature of the serie. For instance, we way try to identify long terms that occur little times. The Fig. 5 is a table of terms longer than 12 characters that occur less than 20 times. Other statistics can be based on other criterion: for example, Fig. 7 is a distribution of the dates printed on the sheets. This includes numbers that are not dates, but the figure shows a significant concentration of these numbers that can be considered as a date. Sequences of 4 digits corresponds often to dates since most of them are concentrated in the period from 1800 to 2000. There are also some interesting peaks that seem to be large collections of a particular period or year.

It is interesting also to perform various statistics on terms to identify a set of rare scientific terms or recurent errors (see Fig. 8). These results may provide a lot of information as long as the number of occurrence is multiple.

G. OCR limitations

It is clear form these results that the quality of OCR can become a real limitation for these kind of analysis [10]. Approximations in OCR analysis can be illustrated by looking at a particular frequent term that generates many mistakes.

For instance, the term vernaculaire occurs 43454 times and it plural version occurs 1997 times (see Fig. 6) The number of occurence of the meanigful term is represented in blue bars on the left side. The red curve represents the level of noise (i.e. terms with a single occurrence are counted on the right side). However, there are many occurrences (727 times) of the term vemaculaire and 720 other version of this term with less than 10 occurrences that have few different letters (vetnaculaiie, vetnaculailc, vetnaculaire, vetnaculaite, veumculaire, etc.). In total, there are 4698 terms which are similar to ”vernaculaire” but have not been detected correctly by the OCR tool.

Another kind of approximation is the different terms used by scientists to call a specimen. Even if the latin term is used, there are several latin terms used for one particular specimen. For instance, the lobelia may be called lobeliaceae (¿166), lobelia (¿768), or lobelianum (4). Finally, terms can be organized in order to extract the level of pertinence according to various criteria, as seen in Fig. 9, even little significant terms may appear in an analysis cloud.

III. The Europeana Newspapers use-case

Europeana is Europe’s digital platform for cultural heritage, providing access to over 54 million of digitized cultural resources from over 3700 cultural heritage institutions, ranging from books, photos and paintings to television broadcasts and 3D objects. It seeks to enable users to search and access knowledge in all the languages of Europe. This is done either directly, via its web portals, or indirectly, via third-party applications built on top of its data services (search APIs and Linked Open Data).

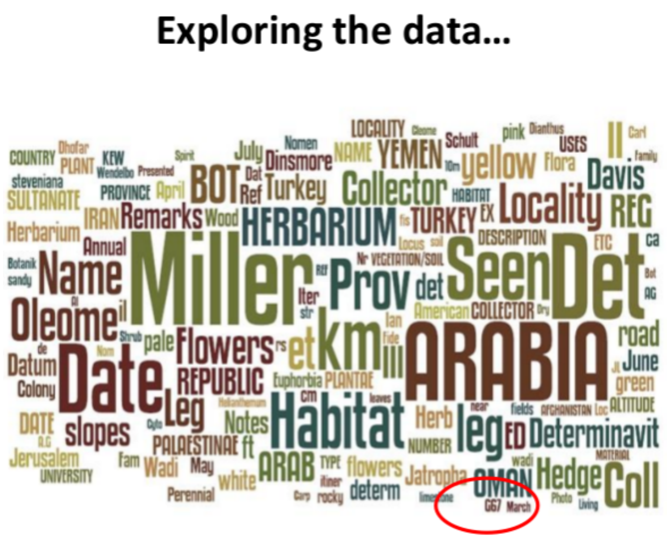

Exploring the data...

Fig. 9. Exploring the data (presented in [7]).

The institutions (libraries, archives, museums, audiovisual collections across Europe) that contribute metadata to Europeana, use a wide range of standards and practices for data and digital object content, and their data arrives to Europeana in dozens of languages. The Europeana service is based on the aggregation and exploitation of data about the digitized objects from very different contexts. To provide a seamless, efficient services on top of such aggregation, it addressed challenging data integration issues. To address these, it necessitates a whole community effort and standardization of data and interoperability across systems. As a hub of cultural heritage data resources, Europeana has developed systems and workflows for ingesting, indexing, normalising and publishing data.

One of Europeanas most recent lines of action is to facilitate the research on the aggregated data resources, especially for the digital humanities and the social sciences. This work is conducted in the scope of Europeana Research, where issues affecting the research re-use of cultural heritage data and content (such as licensing, interoperability and access) are addressed. The earliest work on Europeana Research has clearly identified that, in order to effectively enable the research use of its data resources, it needs to support new use cases face system interoperability challenges. Europeana is currently interested in investigating if, and how, research data e-infrastructures can support its mission to address the requirements for research use of its data resources. The vision is that by leveraging on other European level e-infrastructures for research data, it will be able to reach all potentially interested researchers from all scientific disciplines. Without generic and cross discipline data e-infrastructures, such as

EUDAT, Europeana would have to work with several other e-infrastructures or provide its own research data infrastructure.

With this objective in mind, Europeana is conducting a data pilot with EUDAT. In the datapilot, Europeana is using, as case study, the Europeana Newspapers corpus - one of its datasets that has attracted the most interest from researchers. The pilot investigates the use of the EUDAT services for sharing the corpus through the right mechanisms for its use in research.

A. The Europeana Newspapers corpus

The Europeana Newspapers corpus has been aggregated under a project undertaken in collaboration with several cultural heritage institutions that were active in the digitization of their historic newspapers holdings. It addressed several interoperability aspects of sharing this particular type of digitized materials: metadata descriptions, high resolution images, and full-text obtained from optical character recognition (OCR).

The corpus is nowadays hosted by Europeana, and it comprises, a total of 20 million pages, from which, around 11 million pages were aggregated with full-text. The corpus was made searchable and accessible through a web portal, which attracted much interest from the research community, given it wide variety of content. The corpus was aggregated in Europeana from 12 national and major research libraries in Europe and covers four centuries (1618-2016), four alphabets (Latin, Gothic, Cyrillic and Hebrew) and 40 languages (including historical spelling variants).

B. The Europeana Newspapers datapilot with EUDAT

The collaboration with EUDAT consists in a case study that focus on the Europeana Newspapers corpus. In this section we report on first results obtained from making the two infrastructures interoperable.

1) Objectives of Europeana: The general objective for the datapilot is to investigate how to facilitate the re-use of cultural heritage language resources for research purposes, by using exploiting the EUDAT services. The questions laid out at the start of the datapilot by Europeana were the following:

- How can the resources be discovered
- How can the resources be shared in practical ways for researchers
- How can advanced computation be applied to these Cultural Heritage datasets
- How can the resources and datasets be cited and referenced in research
- How can the Cultural Heritage institutions re-use the outcomes of research on the corpus

2) Evaluation of the EUDAT services for the Europeana Newspaper corpus: Three main areas of work were undertaken in this phase of the data pilot. The following table presents the EUDAT services that have been used, or analyzed in terms of the provided functionality and technical requirements for interoperability between Europeanas systems and EUDAT services.

Area	EUDAT Services	
Publication of cultural heritage datasets	B2SHARE	a service for researchers, scientific communities and citizen scientists to store and share research data from diverse contexts.
	B2SAFE	a robust, safe and highly available service which allows community and departmental repositories to implement data management policies on their research data across multiple administrative domains in a trustworthy manner.
Metadata based discovery of the cultural heritage datasets	B2FIND	a discovery service based on metadata from research data collections from EUDAT data centres and other repositories. The service allows to discover data that is stored through the B2SAFE and B2SHARE services.
Semantic data interoperability between Europeana and EUDAT	B2NOTE	a service that allows the creation of annotations on research data hosted in EUDAT, supporting three types of annotations: semantic tags coming from identified ontology repositories; free-text keywords to be used when a specific semantic term is not found; and free-text comments. B2NOTE is still at the pilot stage. is integrated with the B2SHARE service.

3) Publication of cultural heritage datasets: The publication of the Europeana©Newspapers corpus started with the analysis of the functionality and requirements of B2SHARE and B2SAFE. Only with B2SHARE it would be possible to accomplish the dataset publication requirements of Europeana, although with some scalability concerns due to to the large size of the corpus. A prototype was developed for an interfacing software component that enables the systematic publication of datasets from Europeana into B2SHARE, using the B2SHARE HTTP API. Due to the initial concerns with the scalability of B2SHARE the prototype was tested at the functional level and also with load tests on a dedicated B2SHARE instance deployed for this purpose. Positive and negative results have been obtained after two rounds of testing. On the positive side, the interconnection between the two infrastructures was successfully implemented and functionally tested. It achieved the expectations for making the corpus available for researchers, and for computational processing. The persistent identification supported by EUDAT also fulfilled the requirements for citability. On the negative side, the B2SHARE service has not yet been able to prove that it can scale to the dimension of the Europeana Newspapers corpus. The last round of testing still resulted in many failures due to the overload of B2SHARE. Analysis of the test results by the B2SHARE team has signaled that the test deployment of B2SHARE used in the test was undersized in terms of computational resources.

4) Metadata based discovery of the cultural heritage datasets: For this area of work, the activities focused on determining the data harvesting mechanisms, data formats, and structure of the metadata for making datasets from Europeana discoverable through B2FIND. The ideal harvesting mechanism employed by EUDAT is OAI-PMH, which is the same mechanism used by Europeana to aggregate metadata from its data providers. Basic interoperability for the metadata was achieved with the use of Dublin Core. This basic interoperability results in limitations on the discoverability of Europeanas datasets, therefore, the use of a richer data model is desirable. Future work should address the conversion of the Europeana Data Model, in use at Europeana, into a form fully suitable for B2FIND. Another discoverability issue identified regards the selection of datasets from cultural heritage that are relevant for research purposes. This issue does not affect the Europeana Newspapers corpus, but will affect other datasets from Europeana. Currently, Europeana does not have a definition of criteria for selecting its subsets for research use. A third discovery issue has been identified - the granularity of the item descriptions in the metadata. In most cases, the item level descriptions done in cultural heritage is too fine grained for discovery purposes in research data infrastructures. Europeana and EUDAT will need to work together in both defining criteria for clustering objects and how to exchange metadata about these clusters.

5) Semantic data interoperability between Europeana and EUDAT: Regarding semantic data interoperability between Europeana and EUDAT, we compared the underlying technical solutions at the annotation services of both infrastructures - B2NOTE in EUDAT and Europeanas Annotation API. Also, an assessment was made on the possibilities for semantic interoperability of data from cultural heritage with other scientific disciplines. Both infrastructures employ their specific APIs for annotations, but interoperability between the two services is not be very far away since both services use the same model for representation of annotations - the W3C Web Annotation Data Model. The key area for future work towards semantic interoperability is in the use vocabularies, where It was identified that for reuse of semantic resources from cultural heritage in EUDAT, Europeanas knowledge graph would be the key resource. Europeanas knowledge graph is built from the entities used by Europeanas data providers in the description of the cultural heritage resources that Europeana aggregates. We expect that future activities in this area will be undertaken under the Research Data Alliance Vocabulary Services Interest Group.

IV. Conclusion and future work

In this paper we have shown how to build new knowledge from distributed scientific corpus in the domaine of natural and cultural heritage thanks to active European initiatives EUDAT, Herbadrop and Europeana.

Most results have been produced in an efficient manner since data has been made accessible on the EUDAT Collaborative Data Infrastructure.

For the *Herbadrop* case study, we may assume that, even with a poor quality OCR, it is possible to build substancial knowledge, but as long as the mass of data is important. However, this can become a real limitation for accurate and exhaustive analysis. For this reason, the priority is to investigate in finding better performing OCR tools or more advanced image analysis tools.

As a final assessment, the *Europeana Newspapers* case study has provided a valuable experience for making cultural heritage datasets discoverable, accessible and citable in research contexts. It has identified many requirements for interoperability on both infrastructures: cultural heritage data has some specific characteristics that EUDAT had not encountered before in other communities; and Europeana is identifying the requirements for research usage which may have an impact on how it aggregates data and also in its network of data providers.

In the context of E-Infra12, the goal of the data pilots is to validate the embryo of the future long-term European Trusted Digital Repository (ETDR). The objective of the ETDR is to offer a pan-european service for data preservation and curation into a common infrastructure that would be accessible for non-profit institutions.

V. Acknoledgements

The work done in the Herbadrop data pilot has been made possible thanks to the EUDAT project and the intense collaboration between four European museums and botanical gardens and the CINES. Particular acknoledgements to whom participated directly to Herbadrop: Simon Chagnoux (Musum National d'Histoire Naturelle), Dominik Roepert, Anton Guentsch, Agnes Kirchhoff (Botanischer Garten und Botanisches Museum Berlin), Rob Cubey, Elspeth Haston (Royal Botanic Garden of Edimburgh), Zhengzhe Wu, Hannu Saarenmaa, Ritta Tegelberg (Digitarium, Helsinki), Jean-Christophe Penalva, Alexis Jean-Laurent, Lorene Bechard, Marion Massol, Alexandre Granier, David Rolland, Alexia De Casanove, Audrey Realini, Lucas Guibourt, Samuel Viscapi and special thanks to Benjamin Watine (CINES).

The work done in the Europeana data pilot is partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and by the European Commission under the Connecting Europe Facility, telecommunications sector, grant agreement number CEF-TC-2015-1-01.

References

[1] Ang Y, Puniamoorthy J, Pont AC, Bartak M, Blanckenhorn WU, Eberhard WG, Puniamoorthy N, Silva VC, Munari L, Meier R . A plea for digital reference collections and other science-based digitisation initiatives in taxonomy: Sepsidnet as exemplar. *Systematic entomology*, 38(3):637–644, 2013.

[2] Beaman R, Cellinese N. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. , 209:7–17, 2012.

[3] Berendsohn WG, Guntsch A. Creating a cross-domain pipeline for natural history data. *ZooKeys*, 209:47–52, 2012. http://www.pensoft.net/J_FILES/1/articles/3179/3179-G-3-layout.pdf.

[4] Bunakov Vasily, De Casanove Alexia, Dugenie Pascal, van Horik Rene, Lambert Simon, Quinteros Javier, Reijnhoudt Linda. Data curation policies for EUDAT collaborative data infrastructure. In *Proceedings of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID)*, Moscow, Russia, October 10-13, 2017.

[5] Chagnoux Simon. Herbadrop - 15 months using a pilot infrastructure. In *User Workshop*, Montpellier, Fr., 2017.

[6] Conrad Mark. Collaboration is the thing. In *IEEE Big Data 2016: 1st CAS workshop*, Washington, D.C., USA, 2016.

[7] Cubey Rob. Herbadrop - feedback from the royal botanic garden of edimburgh. In *User Workshop*, Montpellier, Fr., 2017.

[8] Drinkwater RE, Cubey RWN, Haston EM. The use of optical character recognition (OCR) in the digitization of herbarium specimens labels. *PhytoKeys*, 38:15–30, 2014.

[9] DSA. Data Seal of Approval. http://datasealofapproval.org/en/.

[10] Dugenie Pascal, Bechard Lorene. Eudat data pilot herbadrop, final report. 2017.

[11] Dugenie Pascal, Chagnoux Simon. Eudat data pilot herbadrop, second interim report. 2016.

[12] EUDAT. Collaborative Data Infrastructure. https://www.eudat.eu/eudat-cdi.

[13] Haston E, Cubey R, Pullan M, Atkins H, Harris DJ. Developing integrated workflows for the digitization of herbarium specimens using a modular and scalable approach. *ZooKeys*, 209:93–102, 2012.

[14] Haston Elspeth, Chagnoux Simon, Dugenie Pascal. Herbadrop - a long-term preservation of herbarium specimen images. In *Proceedings of the second Eudat User Forum*, Rome, It., 2016.

[15] Johnson Valerie and Ranade Sonia and Thomas David. Size matters: The implications of volume for the digital archive of tomorrow a case study from the uk national archives. *Records Management Journal*, 24(3):224–237, 2014.

[16] Lehtonen J, Heiska S, Pajari M, Tegelberg R, Saarenmaa H. The process of digitizing natural history collection specimens at Digitarium. In *In: Jones MB, Gries C (Eds) Proceedings of the Environmental Information Management Conference 2011 (EIM)*, September 28-29,2011.

[17] OAIS. Open Archival Information System, ISO 14721. https://www.iso.org/standard/57284.html.

[18] RDA. The Research Data Alliance Practical Policy Working Group. https://www.rd-alliance.org/groups/practical-policy-wg.html.

[19] Sandusky Robert J. Computational provenance: Dataone and implications for cultural heritage institutions. In *IEEE Big Data 2016: 1st CAS workshop*, Washington, D.C., USA, 2016.

[20] Simon Hengchen, Mathias Coeckelbergs, Seth van Hooland, Ruben Verborgh, Thomas Steiner. Exploring archives with probabilistic models: Topic modelling for the valorisation of digitised archives of the european commission. In *IEEE Big Data 2016: 1st CAS workshop*, Washington, D.C., USA, 2016.

[21] Ranade Sonia. Traces through time: A probabilistic approach to connected archival data. In *IEEE Big Data 2016: 1st CAS workshop*, Washington, D.C., USA, 2016.

[22] Tegelberg R, Haapala J, Mononen T, Pajari M, Saarenmaa H. The development of a digitising service centre for natural history collections. *ZooKeys*, 209:75–86, 2012.

[23] Tesseract. OCR analysis tool. https://github.com/tesseract-ocr.

[24] The Herbonauts website:. Recruiting the general public to acquire the data from herbarium labels. In *Proceedings of the UNESCO International Conference, Botanists of the twenty-first century: roles, challenges and opportunities* , http://unesdoc.unesco.org/images/0024/002437/243791m.pdf, 22-25 September, 2014.

[25] Weijia Xu, Ruizhu Huang, Maria Esteva, Jawon Song, Ramona Walls. Content-based Comparison for Collections Identification. In *IEEE Big Data 2016: 1st CAS workshop*, Washington, D.C., USA, 2016.